第5章

システム辞書仕様

本章では，EDR から提供されるテキスト形式の辞書（以下，テキスト辞書と呼ぶ）と，辞書トランスレータにより作成される EDBroW 用の辞書（以下，システム辞書と呼ぶ）について，その対応関係を示す．

5.1 各辞書に共通な事柄

概念識別子

テキスト辞書では，概念識別子はそのほとんどが16進数6桁の文字列で記述されているが，一部の概念識別子（数字を表す概念）では，概念識別子が16進数で6桁より大きくなることがある．

また，システム辞書では，概念識別子は unsigned long で格納している．このため，テキスト辞書中で16進数6桁を越える概念識別子については，別途用意するテーブルにより変換した後に格納する．

変換の対象となっているのは，以下の６つの概念識別子である．

テキスト辞書中の 概念識別子	システム辞書中の 概念識別子
000103e8	a00001
00012710	a00002
00010f4240	a00003
000105f5e100	a00004
0001e8d4a51000	a00005
00012386f26fc10000	a00006

なお，上記以外で，テキスト辞書中で6桁を越える概念識別子を含むレコードは，エラーとしてシステム辞書には格納されない．

文法コード

品詞コード，左右連接属性コードについては，テキスト辞書中のコード文字列を，本章末に示すテーブルにより，数値としてシステム辞書に格納する．

また，日本語単語辞書，日英対訳辞書において，"JN1;JVE" という品詞は，サ変名詞を表わすものであり，システム辞書では，"JSA" (品詞番号 38) として格納する．

品詞，連接以外の文法コードについては，文字列をそのまま格納する．

以下に，システム辞書に格納されるデータの名称と，対応するテキスト辞書の名称との対応を示す．

5.2 単語辞書

システム辞書名称	テキスト辞書名称	備考
単語見出し *1	不変化部分	
読み *2	カナ表記（日本語） 音節区切り（英語:慣用句以外） 見出し表記（英語:慣用句）	
発音情報 *3	発音情報（慣用句以外）	
構成語情報 *3	単語見出し（慣用句のみ）	
概念識別子	概念識別子	
左連接情報	左連接情報	慣用句の場合には
右連接情報	右連接情報	格納しない
品詞	品詞	
言語識別 *4	—	
レコード種別 *5	—	
慣用句フラグ *6	—	
文法情報	品詞／構文木以外の 文法情報＋用法	各文法情報コードを ';' で連結して格納
構文木	構文木	慣用句以外は空

*1 慣用句の場合には，先頭の１単語のみを格納する．

*2 英語単語の慣用句の場合，テキスト辞書では音節区切りのフィールドが空であるため，かわりに見出し表記を格納する．

*3 発音情報／構成語情報 は，慣用句とそれ以外のレコードでフィールドが重複しないため，システム辞書中では同フィールドとして格納する．

*4 レコードIDの別により判定．"JWD" を含む場合には 0 を，"EWD" を含む場合には 1 を格納

*5 レコードIDの別により判定．先頭が 'T' で始まる場合には 1 を，それ以外は 0 を格納

*6 主に管理のための情報．レコードが構文木を持つ場合に 1 を，そうでない場合に 0 を格納

※ただし，テキスト辞書の頻度情報は，システム辞書には格納しない

5.3 対訳辞書

システム辞書名称	テキスト辞書名称	備考
単語見出し	単語見出し	
読み	カナ表記（日英）	英日の場合は空
品詞	品詞	
概念識別子	概念識別子	
言語対識別 *1	—	
レコード種別 *2	—	
対訳情報	対訳情報	

*1 レコードIDの別により判定．"JEB" を含む場合には 0 を，"EJB" を含む場合には 1 を格納

*2 レコードIDの別により判定．先頭が 'T' で始まる場合には 1 を，それ以外は 0 を格納

5.4 共起辞書

システム辞書名称	テキスト辞書名称	備考
受け側単語見出し	受け側単語表記	
受け側表記情報	表記情報 *1	カナ表記もしくは原形
受け側品詞	品詞 *1	
受け側概念識別子	概念識別子 *2	
受け側慣用句フラグ	慣用句フラグ *1	
共起関係子	共起関係子	
関係単語表記	関係単語表記	
係り側単語見出し	係り側単語表記	
係り側表記情報	表記情報 *1	カナ表記もしくは原形
係り側品詞	品詞 *1	
係り側概念識別子	概念識別子 *2	
係り側慣用句フラグ	慣用句フラグ *1	
概念関係子	概念関係子 *3	
関係子方向 *4	—	
表記順 *4	—	
言語識別 *5	—	
表層共起頻度	表層共起頻度	
共起項目頻度	共起項目頻度	
受け側共起要素頻度	受け側共起要素頻度	
係り側共起要素頻度	係り側共起要素頻度	
例文	例文	情報を省略して格納 *6

*1 受け側／係り側の各情報は，共起句要素情報の中から，受け側／係り側に対応するもののみを格納

*2 概念識別子が「補足付き概念説明」である場合には，他の概念識別子と重ならない識別子を新設して格納することが可能．ただし，新設した概念識別子については，「補足付き概念説明」文字列を「日本語概念説明」のフィールドに格納した形式の概念見出し補助ファイルを作成し，他の概念見出し辞書ファイルとマージする（辞書トランスレータにより，格納方法を選択可能）

*3 概念関係子が "" である場合には，関係子 "nil"(関係子番号 80) を格納

*4 要素番号をすべて格納しないかわりに，意味情報における始点／終点側と，受け側／係り側との対応をとるために関係子方向を利用．始点＝受け側の場合に 0 を，始点＝係り側の場合に 1 を格納．また，句見出しを格納しないかわりに，句見出しにおける受け側／係り側単語の出現順を表記順とし

て格納する．受け側が先行する場合に 1 を，係り側が先行する場合に 0 を格納.

*5 レコードIDの別により判定．"JCC" を含む場合には 0 を，"ECC" を含む場合には 1 を格納

*6 例文は，１例文につき文IDが複数個ある場合には，先頭の文IDのみを格納し，最大５例文を格納する．

※句見出しは格納しないが，受け側／係り側単語見出し，関係単語表記，表記順により，再構成が可能である．

5.5 概念見出し辞書

システム辞書名称	テキスト辞書名称	備考
概念識別子	概念識別子	
日本語概念見出し	日本語概念見出し	
英語概念見出し	英語概念見出し	
日本語概念説明	日本語概念説明	
英語概念説明	英語概念説明	

5.6 概念体系辞書

システム辞書名称	テキスト辞書名称	備考
上位概念	上位概念	
下位概念	下位概念	

5.7 概念記述辞書

システム辞書名称	テキスト辞書名称	備考
概念識別子１	概念識別子１	
概念識別子２	概念識別子２	
関係子	関係子	
真偽値	真偽値	
記述区分 *1	記述区分	

*1 I記述を 1，E記述を 0 として格納

注：日本語コーパスについて

コーパスのシステムは，他の辞書と形式が異なります.

UNIX上で動作するツールとして「ＥＤＲ辞書管理システム」が提供されています．そのシステムのコーパス部分をＰＣに移植したものです（ホームページにある「辞書利用支援ツール」を参照のこと）.